SRC TR 87-190

Stochastic Approximations for Finite-State Markov Chains

by

D.-J. Ma, A.M. Makowski, and A. Shwartz

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1987		**00-00-1987 to 00-00-1987**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Stochastic Approximations for Finite - State Markov Chains	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Electrical Engineering Department,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**18**	

STOCHASTIC APPROXIMATIONS FOR FINITE - STATE MARKOV CHAINS

by

D.-J. Ma[1], A. M. Makowski[1] and A. Shwartz[2]

University of Maryland and Technion

ABSTRACT

In constrained Markov decision problems, optimal policies are often found to depend on quantities which are not readily available due either to insufficient knowledge of the model parameters or to computational difficulties. This motivates the on-line estimation (or computation) problem investigated in this paper in the context of a single parameter family of finite-state Markov chains. The computation is implemented through an algorithm of the Stochastic Approximations type which recursively generates on-line estimates for the unknown value. A useful methodology is outlined for investigating the strong consistency of the algorithm and the proof is carried out under a set of simplifying assumptions in order to illustrate the key ideas unencumbered with technical details. An application to constrained Markov decision processes is briefly discussed.

Keywords: Markov chains, Recursive Estimation, Stochastic Approximation, Regularity, Implementation, Adaptive Control.

---

[1] Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742, U.S.A. The work of these authors was supported partially through NSF Grant ECS-83-51836, partially through a grant from AT&T Bell Laboratories and partially through ONR Grant N00014-84-K-0614.

[2] Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. The work of this author was supported partially through a Grant from GM Research Laboratories and partially through United States—Israel Binational Science Foundation Grant BSF 85-00306.

1. INTRODUCTION

It is well known that many questions concerning Markov decision processes (MDP's) can be reduced to a search for Markov stationary policies which satisfy certain constraints (or optimality) conditions. However, the authors argued in [5] that the resulting Markov stationary policies are usually *not* readily *implementable*, sometimes in spite of strong structural properties. This is so because the values of the model parameters may not be available [4,9], and even if they were available, the policy may still not be implementable due to computational difficulties inherent to its definition [9].

In this paper, the discussion is given in the context of finite-state MDP's. It is assumed that the policy g of interest belongs to a one-parameter family of Markov stationary policies $\{f^\eta\ , 0 \leq \eta \leq 1\}$ and that the parameter value η^* characterizing g is specified by $J(f^\eta) = V$ for some given scalar V, where $J(\gamma)$ is the cost incurred by using an admissible policy γ. The problem of interest is the *on-line* estimation (or computation) of the parameter η^*, and is solved here through an adaptive algorithm of the Stochastic Approximation type. The adaptive policy α defined through this estimation algorithm is shown to incur the same cost as the policy g, i.e., $J(\alpha) = J(g)$, thus simultaneously resolving the above-mentioned implementation difficulties.

This problem is motivated in Section 3 via an example from the theorey of constrained MDP's, which provides the intuition behind the proposed adaptive algorithm. The convergence results for the estimates and for the cost under the adaptive policy α are presented in Section 4. The method of proof uses the ODE method as discussed by Metivier and Priouret [7], but the specific structure of the model at hand allows for great simplifications in their arguments. The required regularity properties are derived in Section 5 under minimal conditions on the transition probabilities, and the main estimate that underlies the use of the ODE method is developed in Section 6. Section 7 concludes with an application to constrained MDP's and an extension of the results to models with weaker regularity properties.

A few words on the notation used throughout the paper: The set of all real numbers is denoted by $\mathbb{R}$, and $I(A)$ stands for the indicator function of a set A. Unless stated otherwise, $\lim_n$, $\underline{\lim}_n$ and $\overline{\lim}_n$ are taken with n going to infinity.

2. MODEL AND ASSUMPTIONS:

Assume the state space to be a *finite* set S of cardinality d and let the control space U be an arbitrary *measurable* space. The one-step transition mechanism P is defined through the one-step transition probability functions $p_{xy}(\cdot) : U \to \mathbb{R}$ which are assumed to be *Borel* measurable and to satisfy the standard properties $0 \leq p_{xy}(u) \leq 1$ and $\sum_y p_{xy}(u) = 1$ for all x and y in S, and all u in U. The space of probability measures on U (when equipped with its natural Borel σ-field) is denoted by $\mathbb{M}$.

The sample space $\Omega := S \times (U \times S)^\infty$ is the *canonical* space for the MDP (S, U, P). The *coordinate* mappings $\{U_n\}_0^\infty$ and $\{X_n\}_0^\infty$ are defined by setting $U_n(\omega) := u_n$ and $X_n(\omega) := x_n$ for all $n = 0, 1, \ldots$. The sample space Ω is equipped with the σ-field $\mathbb{F} := \vee_{n=0}^\infty \mathbb{F}_n$ where $\mathbb{F}_n := \sigma\{X_0, U_0, X_1, \ldots, U_{n-1}, X_n\}$ for all $n = 0, 1, \ldots$, so that the mappings $\{U_n\}_0^\infty$ and $\{X_n\}_0^\infty$ are all random variables (RV).

An *admissible* control policy γ is defined as any collection $\{\gamma_n\}_0^\infty$ of conditional distributions on U, i.e. , for all $n = 0, 1, \ldots$, the RV $\omega \to \gamma_n(A, \omega)$ is $\mathbb{F}_n$-measurable for every Borel subset A of U, with the interpretation that $\gamma_n(\cdot, \omega)$ is the probability distribution for selecting the control value U_n given the feedback information $(X_0(\omega), U_0(\omega), X_1(\omega), \ldots, U_{n-1}(\omega), X_n(\omega))$. Denote the collection of all such admissible policies by Π.

Let μ be a fixed probability distribution on S. For every admissible policy γ in Π, the Kolmogorov Extension Theorem then guarantees the existence (and uniqueness) of a probability measure P^γ on the σ-field $\mathbb{F}$ so that under P^γ, the RV X_0 has distribution μ and

$$P^\gamma[X_{n+1} = y \mid \mathbb{F}_n] = \int_U \gamma_n(du) p_{X_n y}(u) \qquad n = 0, 1, \ldots \quad (2.1)$$

for all y in S. The expectation operator associated with γ is denoted by E^γ.

A policy γ in Π is said to be a *Markov* or *memoryless* policy if there exists a family $\{g_n\}_0^\infty$ of mappings $g_n : S \to \mathbb{M}$ such that $\gamma_n(\cdot) = g_n(\cdot, X_n)$ $P^\gamma - a.s.$ for all $n = 0, 1, \ldots$ In the event the mappings $\{g_n\}_0^\infty$ are all identical to a given mapping $g : S \to \mathbb{M}$, the Markov policy is termed *stationary* and will be identified with the mapping g itself. For any Markov stationary g, define the $d \times d$ matrix $P(g) = (p_{xy}(g))$ by posing

$$p_{xy}(g) := \int_U p_{xy}(u) g(du, x) \quad (2.2)$$

for all x and y in S.

3. THE IMPLEMENTATION PROBLEM — AN EXAMPLE

For any mapping $c : S \to \mathbb{R}$, define the corresponding long-run average cost functional $J_c : \Pi \to \mathbb{R}$ by posing

$$J_c(\gamma) := \underline{\lim}_n \frac{1}{n+1} E^\gamma \left[\sum_{i=0}^{n} c(X_i) \right] \tag{3.1}$$

for every admissible policy γ in Π.

The problem of interest here is to find a Markov stationary policy g such that $J(g) = V$, with V some real constant determined through various design considerations. Consider the situation where there exist two *implementable* Markov stationary policies $\overline{g}$ and $\underline{g}$ such that

$$J_c(\overline{g}) < V < J_c(\underline{g}), \tag{3.2}$$

i.e., the Markov stationary policy $\overline{g}$ (resp. $\underline{g}$) undershoots (resp. overshoots) the requisite performance level V. This situation arises naturally in the solution of constrained MDP's via Lagrange arguments, and is discussed in Section 7. For every η in the unit interval [0,1], the policy f^η obtained by simply randomizing between the two policies $\overline{g}$ and $\underline{g}$ with *bias* η is the Markov stationary policy determined through the mapping $f^\eta : S \to \mathbb{M}$ where

$$f^\eta(\cdot, x) := \eta\, \underline{g}(\cdot, x) + (1 - \eta)\, \overline{g}(\cdot, x) \tag{3.3}$$

for all x in S. Note that for $\eta = 1$ (resp. $\eta = 0$), the randomized policy f^η coincides with $\underline{g}$ (resp. $\overline{g}$). Owing to the condition (3.2), if the mapping $\eta \to J_c(f^\eta)$ is *continuous* on the interval [0,1], then at least one randomized strategy f^{η^*} meets the value V and its corresponding bias value η^* is a solution of the equation

$$J_c(f^\eta) = V, \quad \eta \text{ in } [0, 1], \tag{3.4}$$

whence $g = f^{\eta^*}$ steers (3.1) to the value V.

Solving the (highly) nonlinear equation (3.4) for the bias value η^* is usually a non-trivial task, even in the simplest of situations [8]. The implementation α of the policy g which is

defined below circumvents this difficulty by bypassing a direct solution of the equation (3.4). The proposed implementations $\alpha = \{\alpha_n\}_0^\infty$ has the form

$$\alpha_n(\cdot, H_n) := \eta_n \, \underline{g}(\cdot, X_n) + (1 - \eta_n) \, \overline{g}(\cdot, X_n) \qquad n = 0, 1, \cdots (3.5)$$

where $\{\eta_n\}_0^\infty$ is some sequence of [0,1]-valued RV's which play the role of "estimates" for the bias value η^*.

In many applications, the mapping $\eta \to J_c(f^\eta)$ is monotone, say monotone increasing for sake of definiteness. The search for η^* can then be interpreted as finding the zero of the monotone function $\eta \to J_c(f^\eta) - V$ and this brings to mind ideas from the theory of *Stochastic Approximations*. Here, this circle of ideas suggests generating a sequence of bias values $\{\eta_n\}_0^\infty$ through the recursion

$$\eta_{n+1} = \Big[\eta_n + a_n\big(V - c(X_{n+1})\big)\Big]_0^1 \qquad n = 0, 1, \ldots (3.6)$$

with η_0 given in [0,1]. In (3.6), the notation $[x]_0^1 = 0 \vee (x \wedge 1)$ is used for every x in $\mathbb{R}$, and the sequence of step sizes $\{a_n\}_0^\infty$ satisfies the conditions

$$0 < a_n \downarrow 0, \qquad \sum_{n=0}^{\infty} a_n = \infty, \qquad \sum_{n=0}^{\infty} a_n^2 < \infty. \qquad (3.7)$$

4. THE RESULTS

The purpose of this note is to provide mild conditions under which (i) the estimates $\{\eta_n\}_0^\infty$ of η^* generated through (3.6) are *strongly consistent* under P^α and (ii) the policies g and α achieve the *same* cost. These results, which are discussed in the remainder of the paper, hold for more general situations with (3.3) replaced by a one-parameter family of stationary policies $\{f^\eta; 0 \le \eta \le 1\}$ such that $f^1 = \underline{g}$ and $f^0 = \overline{g}$ satisfy (3.2). Under a monotonicity assumption, the same reasoning leads to the sequence of bias values generated by (3.6) and to an implementation α of g also given by (3.5). This more general formulation is assumed thereafter and the assumptions of interest can now be stated as conditions (C1)-(C3), where

(C1) Under each policy f^η, the RV's $\{X_n\}_0^\infty$ form an *aperiodic* Markov chain with a *single* recurrent class;

(C2) The transition probabilities $\eta \to p_{xy}(f^\eta)$ are *analytic* on $[0,1]$ for all x and y in S.

(C3) The equation

$$J_c(f^\eta) = V, \qquad 0 \le \eta \le 1 \tag{4.1}$$

has a *unique* solution η^*, and for some $\epsilon > 0$,

$$\big[J_c(f^\eta) - V\big](\eta - \eta^*) > 0 \tag{4.2}$$

whenever $\eta \neq \eta^*$ and $|\eta - \eta^*| \le \epsilon$ in $[0,1]$.

Condition (C2) is relaxed somewhat in Section 7. It is clearly satisfied when f^η is given by (3.3) for then $p_{xy}(f^\eta) = \eta p_{xy}(\underline{g}) + (1-\eta)p_{xy}(\overline{g})$ for all x and y in S. The condition (4.2) is tantamount to local monotonicity and in practice, is often verified by establishing some stronger monotonicity property on $\eta \to J_c(f^\eta)$ such as (C3bis) below.

(C3bis) The mapping $[0,1] \to \mathbb{R} : \eta \to J_c(f^\eta)$ is *strictly monotone*, say monotone increasing for sake of definiteness.

If this mapping is monotone *decreasing*, or if the inequality in (4.2) is reversed, then the stochastic approximation algorithm (3.6) is modified by replacing $(V - c(X_{n+1}))$ with $(c(X_{n+1}) - V)$.

That (4.1) has at least one solution follows from (3.2) and from the following result which is contained in the proof of Theorem 5.4.

Lemma 4.1 *Under the assumptions (C1)-(C2), the mapping* $[0,1] \to \mathbb{R} : \eta \to J_c(f^\eta)$ *is analytic on* $[0,1]$.

The main result of this paper can now be stated.

Theorem 4.2 *Assume (3.2) and (3.7) to hold. Under the assumptions (C1)-(C3), the following statements hold true.*

(i): *The sequence of estimates* $\{\eta_n\}_0^\infty$ *is strongly consistent under* P^α, *i.e.*,

$$\lim_n \eta_n = \eta^* \qquad P^\alpha - a.s. \tag{4.3}$$

(ii): *The policies* g *and* α *achieve the same cost, i.e.,*

$$J_c(\alpha) = J_c(g) = V. \tag{4.4}$$

The approach adopted here for establishing the convergence (4.3) uses an ODE argument based on the deterministic lemma of Kushner and Clark [3] as presented by Metivier and Priouret in [7]. The key result for the analysis is probabilistic in nature and is given the next proposition whose proof is delayed till Section 6. To state the result, consider the RV's $\{Y_n\}_0^\infty$ given by

$$Y_n := J_c(f^{\eta_n}) - c(X_{n+1}) \qquad n = 0, 1, \cdots (4.5)$$

and for every $T > 0$, pose

$$m(n,T) := \max\{k > n : \sum_{i=n}^{k-1} a_i \leq T\} \; . \qquad n = 0, 1, \cdots (4.6)$$

Theorem 4.3 *Under the assumptions (C1)-(C2), the convergence*

$$\lim_n \left(\sup_{n \leq k \leq m(n,T)} | \sum_{i=n}^{k} a_i Y_i | \right) = 0 \qquad P^\alpha - a.s. \qquad (4.7)$$

takes place.

Proof of Theorem 4.2. The result (4.4) on the cost follows readily from the parameter convergence (4.3) upon making use of Theorem 3.1 of [10] which provides extensions to an argument originally due to Mandl [6, Thm. 3, p. 46].

As explained by Metivier and Priouret [7], the convergence (4.7) underlines the P^α-a.s. convergence of $\{\eta_n\}_0^\infty$ to η^*. The reader is invited to consult [3,7] for a complete exposition of the arguments which are now briefly summarized: Interpolate the estimate sequence $\{\eta_n\}_0^\infty$, say by piecewise linear functions $[0, \infty) \to \mathbb{R}$ anchored at η_n at $t_n = \sum_{i=0}^{n-1} a_i$, and define a sequence of left shifts $\eta^{(n)}(t) = \eta(t - t_n)$ which bring the "asymptotic part" of $\{\eta_n\}_0^\infty$ back to a neighborhood of the time origin.

Now observe that the recursion (3.6) can be written in the form

$$\eta_{n+1} = \Big[\eta_n + a_n \big[\big(V - J_c(f^{\eta_n}) \big) + Y_n \big] \Big]_0^1 \qquad n = 0, 1, \ldots (4.8)$$

and that from any convergent subsequence $\{\eta^{(m)}(\cdot)\}_0^\infty$ a further convergent subsequence $\{\eta^{(p)}(\cdot)\}_0^\infty$ can then be extracted by standard boundedness and equicontinuity arguments.

It is then easy to see from Theorem 4.3 that its limit $\eta(\cdot)$, and for that matter the limit of *any* convergent subsequent, satisfies the ODE

$$\dot{\eta}(t) = V - J_c(f^{\eta(t)}), \quad t \geq 0\,, \quad \eta(0) \text{ in } [0,1]\,, \tag{4.9}$$

which is *asymptotically stable* with a *unique* stable point η^*, as a consequence of (C3).

A simple shifting argument now implies $\eta(t) = \eta^*$ for all $t \geq 0$ and this completes the proof. These arguments are now standard and are omitted here for sake of brevity. □

5. SOME REGULARITY RESULTS

The proof of the convergence (4.3) is based on the so-called ODE method as presented by Metivier and Priouret [7]. This approach hinges crucially on the fact that several quantities of interest are *Lipschitz* continuous (in the variable η) and it is the purpose of this section to establish the requisite regularity properties in some detail. In what follows, it will be convenient to view any mapping $f : S \to \mathbb{R}$ as a d dimensional vector $(f(x))$ (still denoted by f). Also, let I_d denote the $d \times d$ identity matrix and let 0_d stand for the $1 \times d$ row vector with **zero entries.**

Note first that in the special case of (3.3), condition (C1) follows from a simple condition on $\overline{g}$ and $\underline{g}$.

Lemma 5.1. *Let f^η be given by (3.3). If both Markov chains $P(\overline{g})$ and $P(\underline{g})$ are irreducible (resp. aperiodic), so is each one of the Markov chains $P(f^\eta), 0 \leq \eta \leq 1$.*

Proof. Note that if for some $n = 0, 1, \cdots$ and some pair of states x and y, either $p_{xy}^{(n)}(\overline{g}) > 0$ or $p_{xy}^{(n)}(\underline{g}) > 0$, then $p_{xy}^{(n)}(f^\eta) > 0$ for all $0 < \eta < 1$. The result now follows readily from the definitions of irreducibility and aperiodicity. □

Under (C1), the Markov chain $P(f^\eta)$ is *positive recurrent* for all $0 \leq \eta \leq 1$ (since S is finite) and therefore possesses a *unique* invariant measure $\pi(\eta)$ which is interpreted as a $1 \times d$ row vector $(\pi(\eta, x))$. It is well known that this invariant vector $\pi(\eta)$ is the *unique* solution to the system of equations

$$\pi = \pi P(f^\eta), \quad \pi e_d = 1 \tag{5.1}$$

in the variable $\pi = (\pi(x))$ in $\mathbb{R}^{1 \times d}$ with e_d denoting the $d \times 1$ column vector with all entries equal to unity.

The next Lemma is useful for establishing the required regularity results. Throughout the discussion, the analyticity of a matrix-valued mapping is understood entrywise.

Lemma 5.2 *If the mapping* $[0,1] \to \mathbb{R} : \eta \to \mathbb{R}^{d\times d} : \eta \to A(\eta) = (A_{xy}(\eta))$ *is analytic with the property that the inverse* $A^{-1}(\eta)$ *of* $A(\eta)$ *exists for every* η *in* $[0,1]$*, then the mapping* $[0,1] \to \mathbb{R} : \eta \to \mathbb{R}^{d\times d} : \eta \to A^{-1}(\eta)$ *is analytic on* $[0,1]$.

Proof. By standard results from Linear Algebra, there exist d^2+1 polynomial functions $r_0 : \mathbb{R}^{d^2} \to \mathbb{R}$ and $r_{xy} : \mathbb{R}^{d^2} \to \mathbb{R}$, with x and y ranging in S, in d^2 variables $A = (A_{xy})$ such that

$$A^{-1}(\eta)_{xy} = \frac{r_{xy}(A(\eta))}{r_0(A(\eta))} \tag{5.2}$$

for all x and y in S and all $0 \le \eta \le 1$. Here, these polynomial functions are of degree at most d and the relation $r_0(A(\eta)) = \det A(\eta) \neq 0$ holds for all $0 \le \eta \le 1$.

It now follows from the expression (5.2) that the mapping $\eta \to A^{-1}(\eta)_{xy}$ is *rational* for all x and y in S, thus analytic throughout [0,1] except possibly at a finite number of points where the function may exhibit poles. However, $r_0(A(\eta))$ is analytic in η and has no zero, so that the assumed analyticity of the mapping $\eta \to A(\eta)$ precludes the existence of poles for each one of the mappings $\eta \to A^{-1}(\eta)_{xy}$ for all x and y in S. □

The smoothness of the components of $\pi(\eta)$ can now be investigated.

Lemma 5.3 *Under (C1)-(C2), the mapping* $[0,1] \to \mathbb{R} : \eta \to \pi(\eta, x)$ *is analytic for every* x *in* S.

Proof. The equations (5.1) satisfied by the invariant vector can be rewritten more compactly as

$$\pi Q(\eta) = [0_d \quad 1] \tag{5.3}$$

where $Q(\eta)$ is the $d \times (d+1)$ matrix given by

$$Q(\eta) := [I_d - P(f^\eta) \quad e_d]. \tag{5.4}$$

Consider the $d \times d$ matrix $\tilde{Q}(\eta)$ obtained from $Q(\eta)$ by removing its first column. Since the invariant measure is uniquely determined by (5.1), it is plain that $\pi(\eta)$ is the *unique* solution to the vector equation $\pi\tilde{Q}(\eta) = [0_{d-1} \quad 1]$ with an obvious interpretation for 0_{d-1}.

Consequently $\tilde{Q}(\eta)$ is *invertible* and

$$\pi(\eta) = [0_{d-1} \quad 1]\tilde{Q}(\eta)^{-1} . \tag{5.5}$$

The mapping $\eta \to \tilde{Q}(\eta)$ is clearly analytic on $[0,1]$ due to (C2) and the result readily follows from Lemma 5.2. □

It is worth pointing out that under (C1),

$$\lim_n \frac{1}{n+1} E^{f^\eta} \left[\sum_{i=0}^{n} 1[X_i = x] \right] = \pi(\eta, x) \tag{5.6}$$

for all x in S (independently of the initial distribution), whence

$$J_c(f^\eta) = \lim_n \frac{1}{n+1} E^{f^\eta} \left[\sum_{i=0}^{n} c(X_i) \right] = \sum_x \pi(\eta, x) c(x) \tag{5.7}$$

Of interest here are the *Poisson* equations associated with the cost c under the policies $f^\eta, 0 \leq \eta \leq 1$. More precisely, the mapping $h : S \to \mathbb{R}$ and the constant J (in $\mathbb{R}$) solve the Poisson equation (associated with c) under policy f^η if

$$h(x) + J = c(x) + \sum_y p_{xy}(f^\eta) h(y) \tag{5.8a}$$

for all x in S, or in equivalent matrix form,

$$h + Je_d = c + P(f^\eta)h . \tag{5.8b}$$

It is clear that if the pair (J, h) solves (5.8) so does $(J, h + ae_d)$ for every a in $\mathbb{R}$. Moreover, it is well known that if the pairs (J_1, h_1) and (J_2, h_2) both solve (5.8), then

$$J_1 = J_2 = \lim_n \frac{1}{n+1} E^{f^\eta} \left[\sum_{i=0}^{n} c(X_i) \right] \tag{5.9}$$

and $h_1 - h_2$ is constant on recurrent classes.

As pointed out earlier, the Markov chain $P(f^\eta)$ has a single positive recurrent class under (C1) (for each $0 \leq \eta \leq 1$), in which case the Poisson equation (5.8) has exactly one solution

$(J_c(\eta), h(\eta))$ where $h(\eta) : S \to \mathbb{R}$ is determined up to an additive constant [11, Thm. 4.1]. A particular representative, still denoted $h(\eta)$, is now described. Before giving this definition, it is convenient to observe that

$$J_c(\eta) = \lim_n \frac{1}{n+1} E^\eta \left[\sum_{i=0}^{n} c(X_i) \right] = J_c(f^\eta) \tag{5.10}$$

as a result of (5.9).

Define the stochastic matrix $P^*(f^\eta)$ by

$$P^*(f^\eta) := \lim_n \frac{1}{n+1} \sum_{i=0}^{n} P(f^\eta)^i. \tag{5.11}$$

This limit exists under (C1) by virtue of elementary results in the theory of Markov chains [2]. Since $P(f^\eta)$ has a single recurrent class, it is plain from (5.6) that all the rows of $P^*(f^\eta)$ are identical to $\pi(\eta)$, so that

$$P^*(f^\eta) = e_d \pi(\eta) \tag{5.12}$$

for all $0 \leq \eta \leq 1$.

It is now a simple exercise to see that the eigenvectors of P^* coincide with those of P, and that the matrix $G(\eta) := P(f^\eta) - P^*(f^\eta)$ has spectral radius *strictly less* than unity, whence $I_d - G(\eta)$ is *invertible*. For all $0 \leq \eta \leq 1$, the mapping $h(\eta) : S \to \mathbb{R}$ is now defined by

$$h(\eta) := [I_d - G(\eta)]^{-1}[I_d - P^*(f^\eta)]c. \tag{5.13}$$

Simple algebraic manipulations show that the pair $(J_c(\eta), h(\eta))$ given by (5.10) and (5.13) solves the Poisson equation (5.8), since $J_c(\eta)e_d = e_d\pi(\eta)c = P^*(f^\eta)c$ by virtue of (5.7) and (5.12).

Theorem 5.4 *Under the assumption (C1)-(C2), the solution pair to the Poisson equation (5.8) given by (5.10) and (5.13) is analytic on* $[0,1]$, *i.e., the mappings* $[0,1] \to \mathbb{R} : \eta \to J_c(f^\eta)$ *and* $[0,1] \to \mathbb{R} : \eta \to h(\eta, x)$, *with* x *ranging over* S, *are all analytic.*

Proof. Since S is finite, the analyticity of the mapping $\eta \to J_c(\eta)$ is an immediate consequence of Lemma 5.3 in view of (5.7) and (5.10).

The matrix-valued function $\eta \to P^*(f^\eta)$ is analytic on [0,1] as a result of the representation (5.13) and of Lemma 5.3. It is now plain that the mappings $\eta \to I_d - P^*(f^\eta)$ and $\eta \to I_d - G(\eta)$ are both analytic on $[0,1]$, and the result now follows from Lemma 5.2. □

As a consequence of Theorem 5.4, since S is finite, there exists a positive constant K such that

$$\left|J_c(\eta) - J_c(\tilde{\eta})\right| \leq K|\eta - \tilde{\eta}| \qquad \text{and} \qquad \sup_x|h(\eta,x) - h(\tilde{\eta},x)| \leq K|\eta - \tilde{\eta}| \tag{5.14}$$

for all $0 \leq \eta, \tilde{\eta} \leq 1$.

6. A PROOF OF THEOREM 4.3

This section is devoted to the proof of the a.s. convergence result (4.7). It is plain from Theorem 5.4 that for each x in S, the mapping $\eta \to h(\eta,x)$ is continuous on $[0,1]$, thus bounded and therefore

$$B := \sup_\eta \sup_x \mid h(\eta,x) \mid < \infty \tag{6.1}$$

since S is finite. Moreover, with the simplified notation E^η for the expectation operator E^{f^η}, the Poisson equation (5.8) easily implies that

$$E^\eta[h(\eta,X_{n+1}) \mid \mathbb{F}_n] = h(\eta,X_n) + J_c(\eta) - c(X_n) \qquad n = 0,1,\cdots \tag{6.2}$$

for all $0 \leq \eta \leq 1$, whence

$$\begin{aligned} &\mid E^\eta[h(\eta,X_{n+1}) \mid \mathbb{F}_n] - E^{\tilde{\eta}}[h(\tilde{\eta},X_{n+1}) \mid \mathbb{F}_n] \mid \\ &= \mid h(\eta,X_n) - h(\tilde{\eta},X_n) + J_c(\eta) - J_c(\tilde{\eta}) \mid \ \leq \ 2K \mid \eta - \tilde{\eta} \mid \end{aligned} \qquad n = 0,1,\ldots \tag{6.3}$$

by making use of (5.14).

It follows from (5.8) that

$$\begin{aligned} -Y_n &= c(X_{n+1}) - J_c(\eta_n) \\ &= h(\eta_n,X_{n+1}) - E^{\eta_n}[h(\eta_n,X_{n+2}) \mid \mathbb{F}_{n+1}] \\ &= Z_n^{(1)} + Z_n^{(2)} + Z_n^{(3)} \end{aligned} \qquad n = 0,1,\cdots \tag{6.4}$$

with

$$Z_n^{(1)} := h(\eta_n,X_{n+1}) - E^{\eta_n}[h(\eta_n,X_{n+1}) \mid \mathbb{F}_n] \tag{6.5a}$$

$$Z_n^{(2)} := E^{\eta_n}[h(\eta_n, X_{n+1}) \mid \mathbb{F}_n] - E^{\eta_{n+1}}[h(\eta_{n+1}, X_{n+2}) \mid \mathbb{F}_{n+1}] \tag{6.5b}$$

and

$$Z_n^{(3)} := E^{\eta_{n+1}}[h(\eta_{n+1}, X_{n+2}) \mid \mathbb{F}_{n+1}] - E^{\eta_n}[h(\eta_n, X_{n+2}) \mid \mathbb{F}_{n+1}] \tag{6.5c}$$

for all $n = 0, 1, \cdots$. Define the RV's $\{S_n^{(k)}\}_0^\infty$ for all $k = 1, 2, 3$, by posing

$$S_n^{(k)} = \sum_{i=0}^{n-1} a_i Z_i^{(k)} \qquad n = 1, 2, \cdots \tag{6.6}$$

with $S_0^{(1)} = S_0^{(2)} = S_0^{(3)} = 0$. It now suffices to show that

$$\lim_n \left(\sup_{n \le \ell \le m(n,T)} | \sum_{i=n}^{\ell} a_i Z_i^{(k)} | \right) = 0 \qquad P^\alpha - a.s. \tag{6.7}$$

for all $T > 0$ and all $k = 1, 2, 3$.

It is plain that the RV's $\{Z_n^{(1)}\}_0^\infty$ form a $(P^\alpha, \mathbb{F}_n)$ martingale-difference, whence $\{S_n^{(1)}\}_0^\infty$ is a zero mean $(P^\alpha, \mathbb{F}_n)$-martingale. Routine calculations show that

$$\sup_n E^\alpha[| S_n^{(1)} |^2] = \sup_n E^\alpha \left[\sum_{i=0}^{n-1} a_i^2 \mid Z_i^{(1)} |^2 \right] \le 4B^2 \sum_{i=0}^{\infty} a_i^2 \tag{6.8}$$

upon using (6.1) and (3.7), and the $(P^\alpha, \mathbb{F}_n)$-martingale $\{S_n^{(1)}\}_0^\infty$ is thus uniformly integrable under P^α. By the Martingale Convergence Theorem, the RV's $\{S_n^{(1)}\}_0^\infty$ converge a.s. under P^α (to an a.s. finite limit), in which case they form a Cauchy sequence P^α-a.s. and (6.7) follows for $k = 1$.

To prove (6.7) for $k = 2$, note that for all $0 \le n < \ell$, the relation

$$\begin{aligned} S_{\ell+1}^{(2)} - S_n^{(2)} &= \sum_{i=n}^{\ell} a_i Z_i^{(2)} \\ &= -\sum_{i=n}^{\ell} (a_{i-1} - a_i) E^{\eta_i}[h(\eta_i, X_{i+1}) \mid \mathbb{F}_i] \\ &\quad + a_{n-1} E^{\eta_n}[h(\eta_n, X_{n+1}) \mid \mathbb{F}_n] - a_\ell E^{\eta_{\ell+1}}[h(\eta_{\ell+1}, X_{\ell+2}) \mid \mathbb{F}_{\ell+1}] \end{aligned} \tag{6.9}$$

holds. It is now plain from (6.1) that

$$| S_{\ell+1}^{(2)} - S_n^{(2)} | \leq B \sum_{i=n}^{\ell} (a_{i-1} - a_i) + B(a_{n-1} + a_\ell) \tag{6.10}$$

$$\leq 2Ba_{n-1} \tag{6.11}$$

upon telescoping the terms in the first sum on the right handside of (6.10) and making use of the monotonicity of the weight sequence $\{a_n\}_0^\infty$. The conclusion (6.7) for $k = 2$ is now immediate.

Finally for $k = 3$, note from (6.3) that

$$| Z_n^{(3)} | \leq 2K | \eta_n - \eta_{n+1} | \qquad n = 0, 1, \cdots \tag{6.12}$$

whereas the recursion (3.6) implies

$$| \eta_{n+1} - \eta_n | \leq a_{n+1} | V - c(X_{n+1}) | \leq a_{n+1}\tilde{B} \qquad n = 0, 1, \cdots \tag{6.13}$$

with $\tilde{B} = V + \sup_x | c(x) |$. By combining (6.12) and (6.13), the inequality

$$| Z_n^{(3)} | \leq 2\tilde{B}Ka_{n+1} \qquad n = 0, 1, \cdots \tag{6.14}$$

is seen to hold and since $\{a_n\}_0^\infty$ is decreasing, this yields the bound

$$\sup_{n \leq \ell \leq m(n,T)} \left| \sum_{i=n}^{\ell} a_i Z_i^{(3)} \right| \leq \sum_{i=n}^{m(n,T)} a_i | Z_i^{(3)} | \leq 2\tilde{B}K \sum_{i=n}^{m(n,T)} a_i^2 \leq 2\tilde{B}Ka_n(T + a_n) \; . \tag{6.15}$$

The convergence (6.7) now follows from (3.7). □

7. CONCLUDING REMARKS

The results of this paper can be given the interpretation either of an estimation procedure, where the estimated parameter is defined through (3.4), or of an adaptive implementation scheme, where the controls are generated "on line" through (3.6). The paper concludes with an application to constrained MDP's and with several extensions of the results.

Constrained optimization

The results of Section 4 have an immediate application to the following problem. Let c and d be two cost functions $S \to \mathbb{R}$ and denote the corresponding long-run average costs incurred by an arbitrary policy γ in Π, as defined in (3.1), by $J_c(\gamma)$ and $J_d(\gamma)$, respectively. With $\Pi_V := \{\gamma \in \Pi : J_c(\gamma) \geq V\}$ for some V in $\mathbb{R}$, consider the *constrained optimization* problem

$$\text{Maximize} \quad J_d(\cdot) \quad \text{over} \quad \Pi_V.$$

In the event $c \leq 0$ and $d \geq 0$, the problem has a natural interpretation of maximizing the reward subject to a bound on the cost. Assume henceforth that Π_V is non-empty and strictly contained in Π, so that the problem is feasible but not trivial.

Beutler and Ross [1] have shown that if U is *compact* and if the mappings $u \to p_{xy}(u)$ are continuous for all x and y in S, then there exist two *Markov deterministic* policies $\overline{g}$ and $\underline{g}$ so that (3.2) holds. Moreover, if f^η is given by (3.3), then $\eta \to J_c(f^\eta)$ is continuous, and if η^* solves (3.4), then $g = f^{\eta^*}$ is a solution to the constrained optimization problem.

Applying the results of Theorem 4.2, it follows that if $\eta \to J_c(f^\eta)$ satisfies condition (C3), then the policy α obtain through (3.5)-(3.6) satisfies $J_c(\alpha) = J_c(g) = V$. Similarly, $J_d(\alpha) = J_d(g)$ and α solves the constrained optimization problem.

Extensions

The results of this paper can be obtained under regularity conditions which are much weaker than (C2). One possible set of conditions under which the analysis carries through is stated as condition (C2bis) below, where

(C2bis) The transition probabilities $\eta \to p_{xy}(f^\eta)$ are *Hölder continuous* for all x and y in S, i.e. there exist constants $K > 0$ and $0 < \beta \leq 1$, such that

$$|p_{xy}(f^\eta) - p_{xy}(f^{\tilde{\eta}})| \leq K|\eta - \tilde{\eta}|^\beta \tag{7.1}$$

for all x and y in S.

In exact parallel with the developments of Sections 5 and 6, conditions (C1), (C2bis) and (C3) are sufficient to guarantee that

(i): For all x in S, the mapping $\eta \to \pi(\eta, x)$ is Hölder continuous with parameter β.

(ii): The mappings $\eta \to J_c(f^\eta)$ and $\eta \to h(\eta, x)$, with x ranging over S, are all Hölder continuous with parameter β.

(iii): If $\{\eta_n\}_0^\infty$ is given by (3.6), then (4.3) and (4.4) hold.

The proofs of (i)-(ii) are identical to the ones given for Lemma 5.3 and Theorem 5.4, respectively, upon observing that the class of Hölder continuous functions with parameter β is closed under addition and multiplication, and under composition with the function $x \to \frac{1}{x}$ on closed intervals which do not include 0. The proof of Theorem 4.3 carries over with a slight modification, namely that the last term in (6.3) and (6.12) needs to be changed to $2K|\eta - \tilde{\eta}|^\beta$. Modifying (6.14)-(6.15) appropriately, the last bound in (6.15) becomes $2\tilde{B}^\beta K a_n^\beta (T + a_n)$, which converges to zero due to (3.7).

If the regularity postulated in (C2bis) is changed to continuous differentiability of order r, then the same remarks show that the smoothness in (i)-(ii) will then also be of order r.

REFERENCES

[1] F. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *J. Math. Anal. Appl.*, Vol. 112, pp. 236-252 (1985).

[2] K. L. Chung, *Markov Chains with Stationary Transition Probabiliites*, Second Edition, Springer - Verlag, New York (1967).

[3] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Applied Mathematical Sciences, Vol. 26, Springer-Verlag, Berlin, 1978.

[4] D.-J. Ma and A. M. Makowski, "A simple problem of flow control II: Implementation of threshold policies via Stochastic Approximations," *IEEE Trans. Auto. Control* (submitted 1987).

[5] A. M. Makowski and A. Shwartz, "Implementation issues for Markov decision processes," *Proceedings of a Workshop on Stochastic Differential Systems*, Institute of Mathematics and its Applications, University of Minnesota, Eds. W. Fleming and P.-L. Lions, Springer Verlag Lecture Notes in Control and Information Sciences (1986).

[6] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.*, Vol. 6, pp. 40-60 (1974).

[7] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Info. Theory*, Vol. AC-30, pp. 140-150 (1984).

[8] P. Nain and K. W. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Auto. Control*, Vol. AC-31, pp. 883-888 (1986).

[9] A. Shwartz and A. M. Makowski, "An optimal adaptive scheme for two competing queues with constraints", *Proceedings of the 7th International Conference on Analysis and Optimization of Systems*, Eds. A. Bensoussan and J.-L. Lions, Springer Verlag Lecture Notes in Control and Information Sciences, pp. 515-532, Antibes, France (1986).

[10] A. Shwartz and A. M. Makowski, "Comparing policies in Markov decision processes: Mandl's Lemma revisited," *Mathematics of Operations Research* (submitted 1987).

[11] A. Shwartz and A. M. Makowski, "On the Poisson equation for Markov chains," *Mathematics of Operations Research* (Submitted 1987).